Understanding the Market through Sentiment Analysis

Agam Singh

Data

StockNews:
https://www.kaggle.com/aaron7sun/stocknews

- 1988 rows (25 headlines each and 2 labels)
- 80/20 train test split
- Label 1 for experiment 1 (stock returns)
- Label 2 for experiment 2 (stock volatility)

Experiment (1)

Given the top 25 news headlines on some day, can we predict the directions of next day's market returns?

$$R_t = \ln\left(\frac{S_t}{S_{t-1}}\right)$$

Note: "Direction" of market returns is defined as the sign of the returns.

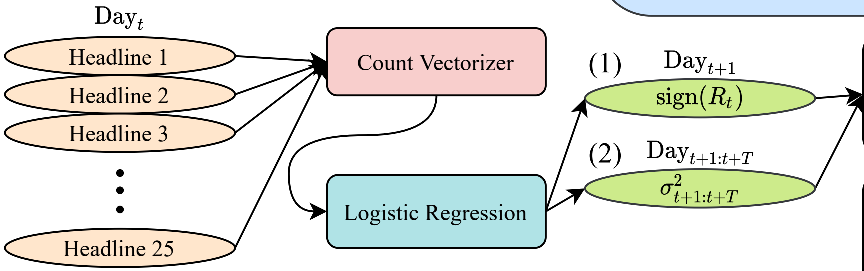
Experiment (2)

Given the top 25 news headlines on some day, can we predict the rolling realized volatility over the next 14 days?

$$\sigma^2_{t-1:t+T} = \frac{252}{T}\sum_{i=1}^{T}\left[\ln\left(\frac{S_i}{S_{i-1}}\right)\right]^2$$

Note: We are forecasting the rolling realized volatility.

A Baseline Model:



Results

Experiment 1 accuracy: 55%
Experiment 2 accuracy: 43%

Why these?

Experiment 1: random variation
Experiment 2: overfitting!

Highest Weighted Words (2)

"oil spill"

"wall street"

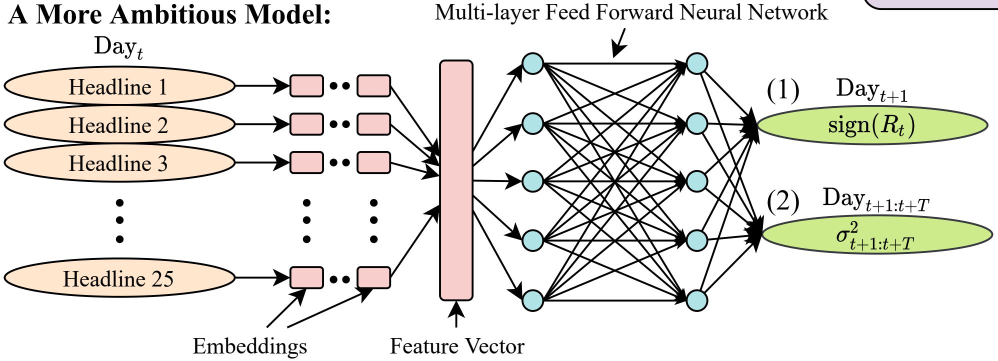
"in pakistan"

"south africa"

"hong kong"

"vladimir putin"

A More Ambitious Model:



Why a neural network? Embeddings?

We hope that we can better learn the nonlinearities present in the data!

More!

Come up with a way to handle current events to remove overfitting!